Encoding method for the compression of a video sequence

The present invention relates to an encoding method for the compression of a video sequence divided into groups of frames and decomposed by means of a three-dimensional (3D) wavelet transform leading to a given number of successive resolution levels that correspond to the decomposition levels of said transform, said method being based on a hierarchical subband encoding process leading from the original set of picture elements (pixels) of each group of frames to transform coefficients constituting a hierarchical pyramid, a spatio-temporal orientation tree - in which the roots are formed with the pixels of the approximation subband resulting from the 3D wavelet transform and the offspring of each of these pixels is formed with the pixels of the higher subbands corresponding to the image volume defined by these root pixels - defining the spatio-temporal relationship inside said hierarchical pyramid, the subbands to be encoded being scanned one after the other in an order that respects the parent-offspring dependencies formed in said tree and preserves the initial subband structure of the 3D wavelet transform.

The video streaming over heterogeneous networks requires a high scalability capability. That means that parts of a bitstream can be decoded without a complete decoding of the sequence and can be combined to reconstruct the initial video information at lower spatial or temporal resolutions (spatial/temporal scalability) or with a lower quality (PSNR scalability). A convenient way to achieve all the three types of scalability is a three-dimensional (3D) wavelet decomposition of the motion compensated video sequence.

In a previous European patent application filed by the Applicant on May 3, 2000, with the number 00401216.7 (PHFR000044), a simple method of texture coding having this property has been described. In that method, as well as in other published documents (such as for instance, in "An embedded wavelet video coder using three-dimensional set partitioning in hierarchical trees (SPIHT)", by B. Kim and W.A. Pearlman, Proceedings DCC'97, Data Compression Conference, Snowbird, UT, USA, 25-27 March 1997, pp.251-260), all the motion vector fields are encoded and sent in the bitstream, which may become a major drawback when a low bitrate is targeted and the receiver only wants a reduced frame rate or spatial resolution.

It is therefore an object of the invention to propose an encoding method more adapted to the situation where a high scalability must be obtained.

To this end, the invention relates to an encoding method such as defined in the introductory part of the description and which is moreover characterized in that, in view of a temporal scalability, a motion estimation is performed at each temporal resolution level, the beginning of which is indicated by flags inserted into the bitstream, and only the estimated motion vectors necessary to reconstruct any given temporal resolution level are encoded and put in the bitstream together with the bits encoding the wavelet coefficients at this given temporal level, said motion vectors being inserted into said bitstream before encoding texture coefficients at the same temporal level.

In another embodiment, the invention also relates to an encoding method such as defined in said introductory part and which is characterized in that, in view of a spatial scalability, a motion estimation is performed at the highest spatial resolution level, the vectors thus obtained being divided by two in order to obtain the motion vectors for the lower spatial resolutions, and only the estimated motion vectors necessary to reconstruct any spatial resolution level are encoded and put in the bitstream together with the bits encoding the wavelet coefficients at this given spatial level, said motion vectors being inserted into said bitstream before encoding texture coefficients at the same spatial level, and said encoding operation being carried out on the motion vectors at the lowest spatial resolution, only refinement bits at each spatial resolution being then put in the bitstream bitplane by bitplane, from one resolution level to the other.

The technical solution thus proposed allows to encode only the motion vectors corresponding to the desired frame rate or spatial resolution, instead of sending all the motion vectors corresponding to all possible frame rates and all spatial resolution levels.

The present invention will now be described, by way of example, with reference to the accompanying drawings in which :

Fig.1 illustrates a temporal subband decomposition of the video information with motion compensation using the Haar multiresolution analysis ;

Fig.2 shows the spatio-temporal subbands resulting from a three-dimensional wavelet decomposition ;

Fig.3 illustrates the motion vector insertion in the bitstream for temporal scalability ;

Fig.4 shows the structure of the bitstream obtained with a temporally driven scanning of the spatio-temporal tree ;

Fig.5 is a binary representation of a motion vector and its progressive transmission from the lowest resolution to the highest ;

Fig.6 shows the bitstream organization for motion vector coding in the proposed scalable approach.

A temporal subband decomposition of a video sequence is shown in Fig.1. The illustrated 3D wavelet decomposition with motion compensation is applied to a group of frames (GOF), referenced F1 to F8. In this 3D subband decomposition scheme, each GOF of the input video is first motion-compensated (MC in Fig.1) (this step allows to process sequences with large motion) and then temporally filtered using Haar wavelets (the dotted arrows correspond to a high-pass temporal filtering, while the other ones correspond to a low-pass temporal filtering) and after these two operations, each temporal subband is spatially decomposed into a spatio-temporal subband, which leads to a 3D wavelet representation of the original GOF, as illustrated in Fig.2. In Fig.1, three stages of decomposition are shown (L and H = first stage ; LL and LH = second stage ; LLL and LLH = third stage). At each temporal decomposition level of the illustrated group of 8 frames, a group of motion vector fields is generated (MV4 at the first level, MV3 at the second one, MV2 at the third one). When a Haar multiresolution analysis is used for the temporal decomposition, since one motion vector field is generated between every two frames in the considered group of frames at each temporal decomposition level, the number of motion vector fields is equal to half the number of frames in the temporal subband, i.e. four at the first level of motion vector fields, two at the second one, and one at the third one. At the decoder side, in order to reconstruct a given temporal level, only the motion vector fields at that level and at the lower temporal resolutions (reduced frame rate) are needed.

(A) Temporal scalability

This observation leads, according to the invention, to organize the bitstream in a way that allows a progressive decoding, as described for example in Fig. 3 : three temporal decomposition levels TDL (as shown in Fig.1) yield four temporal resolution levels (1 to 4), which represent the possible frame rates that can be obtained from the initial frame rate. The coefficients corresponding to the lowest resolution temporal level are first encoded, without sending motion vectors at this level, and, for all the other reconstruction frame rates, the motion vector fields and the frames of the corresponding high frequency temporal subband

are encoded. This description of the bitstream organization up to now only takes into account the temporal levels. However, for a complete scalability, one has to consider the spatial scalability inside each temporal level . The solution for wavelet coefficients was described in the European patent application already cited, and it is reminded in Fig. 4 : inside each temporal scale, all the spatial resolutions are successively scanned (SDL = spatial decomposition levels), and therefore all the spatial frequencies are available (frame rates t = 1 to 4 ; display sizes s = 1 to 4). The upper flags separate two bitplanes, and the lower ones two temporal decomposition levels.

(B) Spatial scalability

In order to be able to reconstruct a reduced spatial resolution video, it is not desirable to transmit at the beginning of the bitstream the motion vector fields of full resolution. Indeed, it is necessary to adapt the motion described by the motion vectors to the size of the current spatial level. Ideally, it would be desirable to have first a low resolution motion vector field corresponding to the lowest spatial resolution and then to be able to progressively increase the resolution of the motion vectors according to the increase in the spatial resolution. Only the difference from a motion vector field resolution to another one would be encoded and transmitted.

It will be assumed that the motion estimation is performed by means of a block-based method like full search block matching or any other derived solution, with an integer pixel precision on full resolution frames (this hypothesis does not reduce the generality of the problem : if one wants to work with half-pixel precision for motion vectors, by multiplying by 2 all the motion vectors at the beginning, one returns in the previous case of integer vectors, even though they will represent fractional displacements). Thus, motion vectors are represented by integers. Given the full resolution motion vector field, in order to satisfy the above requirements of spatial scalability, the motion vector resolution is reduced by a simple divide-by-2 operation. Indeed, as the spatial resolution of the approximation subband is reduced by a factor 2, while the motion is the same as in the full resolution subband, the displacements will be reduced by a factor 2. This division is implemented for integers by a simple shift.

The size of the blocks in the motion estimation must be chosen carefully : indeed, if the original size of the block is 8x8 in the full resolution, it will become 4x4 in the half resolution, then 2x2 in the quarter, and so on. A problem will therefore appear if the original size of the blocks is too small : the size can be null for small spatial resolutions. Thus

it must be checked that the original size is compatible with the number of decomposition/reconstruction levels.

It is now assumed that one has S spatial decomposition levels and that one wants the motion vectors corresponding to all possible resolutions, from the lowest to the highest. Then, either the initial motion vectors are divided by 2^S or a shift of S positions is performed. The result represents the motion vectors corresponding to the blocks from lowest resolution whom the size is divided by 2^S. A division by 2^{S-1} of the original motion vector would provide the next spatial resolution. But this value is already available from the previous operation. Indeed, it corresponds to a shift of S - 1 positions. The difference from the first operation is the bit in the binary representation of the motion vector with a weight of 2^{S-1}. It is then sufficient to add this bit (the refinement bit) to the previously transmitted vector to reconstruct the motion vector at a higher resolution, which is illustrated in Fig. 5 for $S = 4$. This progressive transmission of the motion vectors allows to include in the bitstream the refinement bits of the motion vector fields from one spatial resolution to another just before the bits corresponding to the texture at the same spatial level. The proposed method is resumed in Fig. 6.

The motion vectors at the lowest resolution are encoded with a DPCM technique followed by entropy coding using usual VLC tables (e.g., those used in MPEG-4). For the other resolution levels, a complete bitplane composed of the refinement bits of the motion vector field has to be encoded, for instance by means of a contextual arithmetic encoding, with the context depending on the horizontal or vertical component of the motion vector.

The part of the bitstream representing motion vectors precedes any information concerning the texture. The difference with respect to a "classical" non-scalable approach is that the hierarchy of temporal and spatial levels is transposed to the motion vector coding. The most significant improvement with respect to the previous technique is that the motion information can be decoded progressively. For a given spatial resolution, the decoder does not have to decode parts of the bitstream that are not useful at that level.